Caracterización de series de tiempo con escalamiento multidimensional

V. Rivera-Mancera [1], E. Bautista-Thompson [2], J. Figueroa-Nazuno [3]

Centro de Investigación en Computación, Instituto Politécnico Nacional
Unidad Profesional "Adolfo López Mateos" Zacatenco
Col. Linda Vista, C.P. 07738, México D.F.
rivera_mancera@hotmail.com [1] ebautista@correo.cic.ipn.mx [2]
jfn@cic.ipn.mx [3]

Resumen. Dentro de las técnicas de análisis multivariadas podemos citar el Escalamiento Multidimensional (Multidimensional Scaling MDS). El MDS es una técnica de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones, las proximidades existentes entre un conjunto de objetos en base a métricas de similitud. El problema de identificar series de tiempo con propiedades similares que son indicativas de su dificultad de predicción, motivó en el presente trabajo, el aplicar la técnica MDS para identificar agrupaciones de series que cumplen lo anterior. Para ello se caracterizó un conjunto de 30 series de tiempo de diferentes orígenes (natural y artificial), a las cuales se les calculó un conjunto de propiedades representativas de origen computacional, topológico, estadístico, espacial, y temporal

Introducción

El escalamiento multidimensional (Multidimensional Scaling MDS), permite obtener información cuantitativa y cualitativa de las posibles relaciones de similitud entre objetos (en este caso de series de tiempo), mediante el escalamiento multidimensional métrico y el escalamiento multidimensional no-métrico respectivamente. El MDS es una técnica de representación espacial que trata de visualizar sobre un mapa, un conjunto de propiedades cuya posición relativa se desea analizar. El propósito del MDS es transformar las propiedades de las series de tiempo y llevarlas a una matriz de distancias susceptible de ser representada en un espacio multidimensional. El MDS está basado en la comparación de objetos, de forma tal que si un individuo juzga a los objetos A y B como los más similares, es debido a que la técnica de MDS coloca a los objetos A y B en el gráfico a una distancia entre ellos que sea la menor respecto a cualquier otra distancia entre cualquier otro par de objetos [1, 2].

Existen otras técnicas multivariadas, como son el análisis factorial y el análisis cluster, que persiguen objetivos muy similares al MDS pero difieren en una serie de aspectos, cómo por ejemplo: en el MDS no es necesario especificar cuáles son las variables a emplear en la comparación de objetos, algo que es fundamental en el análisis factorial y en el análisis cluster. Sin embargo, la utilización de alguna de estas técnicas no supone que no se pueda utilizar el escalamiento multidimensional, sino

A. Gelbukh, M. Hernández Cruz (Eds.) Avances en la ciencia de la computación en México, CORE-2003, pp. 171–180, 2003.

que esta última técnica puede servir como complemento a las otras técnicas multivariadas.

Los objetivos de éste trabajo son:

1) Encontrar propiedades que indiquen la predictibilidad de las diferentes series tiempo.
2) Aplicar la técnica MDS para la caracterización de las mismas.

2 El modelo general de escalamiento multidimensional

De modo general, podemos decir que el MDS toma como entrada una matriz proximidades, $\Delta \in M_{nxn}$, donde n es el número de series de tiempo. Cada elemento δ_{ij} de Δ representa la proximidad entre las propiedades para las series de tiempo ST_i y ST_j donde $\delta_{ij} = (P_i - P_j)^2$, P_k es una propiedad de la serie de tiempo k-ésima.

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{bmatrix}$$

El MDS inicia con una matriz aleatoria $X \in M_{nxm}$, donde n, al igual que antes, número de series de tiempo, y m es el número de dimensiones. En nuestro tomaremos $m = 2$. Cada valor x_{ij} representa la coordenada de la ST_i en dimensión j.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

A partir de esta matriz X se puede calcular la distancia existente entre dos series tiempo cualesquiera i y j, simplemente aplicando la fórmula general de la distancia Minkowski:

$$d_{ij} = \left[\sum_{t=1}^{m} \left(x_{it} - x_{jt} \right)^p \right]^{1/p}$$

donde p puede ser un valor entre 1 e infinito, en nuestro caso tomaremos $p =$ partir de estas distancias podemos obtener una matriz de distancias que denominan $D \in M_{nxn}$.

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \tag{4}$$

La solución proporcionada por el MDS debe ser de tal modo que haya la máxima correspondencia entre matriz de proximidades inicial Δ y la matriz de distancias obtenidas D, la cuál se logra modificando la matriz X de forma iterativa.

$$X=\frac{BX}{2n} \tag{5}$$

En donde B tiene como elementos:

$$b_{ij}=\frac{-2\delta_{ij}}{d_{ij}} \quad si \quad i\neq j \tag{6}$$

$$b_{ii}=\sum\sum\frac{2\delta_{ik}}{d_{ik}} \quad si \quad i=j \tag{7}$$

$$b_{ij}=0 \quad si \quad d_{ij}=0 \tag{8}$$

3 Modelos de escalamiento multidimensional

Existen dos modelos básicos de MDS que son: el modelo de escalamiento métrico y el modelo de escalamiento no métrico. En el primero de ellos consideramos que los datos están medidos en escala de razón o en escala de intervalo y en el segundo consideramos que los datos están mediados en escala ordinal. No se ha desarrollado todavía ningún modelo para datos en escala nominal. Para el estudio de las series de tiempo sólo usaremos el modelo de escalamiento no métrico, ya que es éste el que permite formar clases o agrupaciones, a diferencia del escalamiento métrico que arroja información cuantitativa.

Modelo de escalamiento no métrico

A diferencia del escalamiento métrico, el modelo de escalamiento no métrico no presupone una relación lineal entre las proximidades y las distancias, sino que establece una relación monótona creciente entre ambas, es decir, si $\delta_{ij}<\delta_{kl}\Rightarrow d_{ij}\leq d_{kl}$. Su desarrollo se debe a Shepard (1962) quién demostró que es posible obtener soluciones métricas asumiendo únicamente una relación ordinal entre proximidades y distancias. Posteriormente Kruskal (1964) mejoró el modelo [1, 2].

El procedimiento se basa en los siguientes apartados:

1) Obtención de una matriz $X\in M_{nxm}$ de coordenadas aleatorias, que nos da la distancia entre las series de tiempo.

2) Comparación de las proximidades con las distancias contenidas en obteniéndose las disparidades.
3) Cálculo del S-Stress (explicado más adelante en está sección).
4) Modificación de la matriz X mediante la ecuación 5 con el fin de minimizar el Stress.
5) Ir al paso 2 hasta que el S-Stress alcance el valor deseado

Tanto para el modelo métrico como para el modelo no métrico es necesario obtener un coeficiente que nos informe sobre la precisión del modelo. Sabemos que las distancias son una función de las similitudes, es decir:

$$f : \delta_{ij}(x) \rightarrow d_{ij}(x) \tag{9}$$

de esta forma se tiene que $d_{ij} = f(\delta_{ij})$. Esto no deja ningún margen de error. Sin embargo, en las proximidades empíricas es difícil que se dé la igualdad, por lo que generalmente ocurre que $d_{ij} \approx f(\delta_{ij})$, $f(\delta_{ij}) = a + b\delta_{ij}$ donde a y b son constantes que se deben de determinar. A las transformaciones de las proximidades por f se denomina *disparidades.* A partir de aquí podemos definir el error cuadrático como:

$$e_{ij}^2 = (f(\delta_{ij}) - d_{ij})^2$$

Como medida de la precisión del modelo utilizaremos el S-*Stress* definido como:

$$S-Stress = \sqrt{\frac{\sum_{i,j}\left(f(\delta_{ij})^2 - d_{ij}^2\right)^2}{\sum_{i,j}\left(d_{ij}^2\right)^2}} \tag{10}$$

4 Análisis experimental de las series de tiempo

Se tomaron 30 series de tiempo que han sido reportadas en la literatura especializad en la evaluación de técnicas de predicción y modelado [3, 8] y que además corresponden a diversos orígenes (experimental o generadas por modelc matemáticos). La selección de las propiedades calculadas que caracterizan a las serie de tiempo se realizo considerando que cuantifican características de tipo estadístico topológico, computacional, espacial y temporal [3, 4, 7].

A continuación se da una descripción de las propiedades calculadas:

Exponente de Lyapunov. El exponente principal de Lyapunov, mide la evolución trayectorias vecinas en el espacio fase. Mide la inestabilidad de la dinámica sistema debido a cambios en sus condiciones iniciales [3, 6].

Exponente de Hurst. El exponente de Hurst permite determinar si el fenómeno representado por la serie de tiempo presenta correlaciones de largo alcance (memoria y persistencia de largo alcance) [3, 6].

Dimensión de Capacidad. La dimensión de capacidad es similar a la dimensión de Hausdorff y mide el grado de auto-similitud del sistema (comportamiento invariante ante cambios de escala espacial), permite cuantificar el grado de heterogeneidad de la señal a diferentes escalas [3, 6].

Dimensión de Correlación. La dimensión de correlación mide la cantidad de veces que la trayectoria del sistema pasa por una vecindad dada en el espacio fase, cuantifica la correlación espacial local entre puntos de la trayectoria en el espacio fase, sin tomar en cuenta el grado de correlación temporal [3, 5].

Frecuencia Dominante. Permite determinar si existe alguna frecuencia característica de la señal. Las series de carácter aleatorio poseen espectros amplios sin ninguna frecuencia dominante, lo mismo se aplica en cierto grado para las series caóticas. Las series periódicas y cuasi-periódicas poseen picos bien definidos [3, 5].

Entropía Espacio-Temporal. La entropía-temporal, cuantifica de forma global la no correlación de los datos mediante el análisis de recurrencia [3, 5].

Entropía de Shannon. La entropía de Shannon, es una medida de la cantidad de información que se obtiene al tomar una medida para especificar el estado del sistema [5, 6].

Porcentaje de Determinismo. Permite medir el grado de determinismo en el sistema, por medio del análisis de mapas de recurrencia [3, 5].

Porcentaje de Recurrencia. Permite medir el grado de recurrencia (periodicidad y estructura) entre los datos de la serie, que es indicativo de patrones repetitivos en la serie de tiempo, por medio del análisis de mapas de recurrencia [3, 5].

Reglas de Producción. La generación de gramáticas a partir de una serie de tiempo permite dar una medida de complejidad (computacional) en la cual a mayor número de reglas de producción necesarias para generar una gramática, mayor es la dificultad para la predicción o modelado de la serie de tiempo [3, 8].

La tabla 1 nos muestra las 30 series de tiempo y los valores calculados.

5 Método experimental del escalamiento multidimensional

Una vez obtenidas las propiedades de las series de tiempo se aplicó el procedimiento de MDS, para cada una de las propiedades para el conjunto de 30 series de tiempo.

El diagrama 1 muestra el procedimiento de MDS para las series de tiempo, usando el algoritmo ALSCAL (Alternating Least Squares SCALing), en el cuál se usaron las propiedades de las series de tiempo, para construir una matriz a través de la ecuación $\delta_{ij} = (P_i - P_j)^2$. Con esta matriz se cálculo la matriz de distancias. Este proceso se repitió hasta que el S-Stress presento una mejoría mínima de 0.001.

Tabla 1. Propiedades calculadas experimentalmente de las treinta series de tiempo

series de tiempo	exp. Hurst	dim. capacidad	dim. correlación	exp. Lyapunov	frecuencia Dominante	reglas de producción	entropía (Shannon)
Seno	0.948	0.246	0.228	0.517	-0.001	27	3.712
Qp2	0.002	0.679	0.923	0.925	0.13	66	2.446
Mkg	0.078	0.983	1.025	1.481	0.068	75	0.159
Log	-0.059	0.941	0.93	0.76	0.38	61	0.035
Lorentz	0.756	0.965	1.025	0.601	0.001	63	4.789
Rossler	0.562	0.994	1.026	1.049	0.017	72	2.459
Ikeda	-0.02	0.983	1.019	1.452	0.324	71	
Henon	-0.33	0.997	0.991	2.301	0.46	69	0.02
D1	0.312	0.986	2.053	1.288	0	92	0.852
Laser	0.088	0.961	2.096	0.949	0.13	66	2.935
Dowj	0.563	0.887	1.035	0.144	0	65	6.435
Kobe	0.016	0.904	1.053	1.102	0.22	79	3.744
Ecg	0.748	0.934	1.051	0.242	0	53	3.51
Eeg	0.204	0.890	1.876	1.0199	0.041	82	
Brownian	0.537	0.971	1.103	2.043	0	85	2.74
Whiten	0.001	0.983	2.086	1.606	0	73	
Ascii	-0.147	0.485	2.011	2.709	0.288	80	2.107
Cantor	0.001	0.661	0.658	5.728	0.422	72	
El niño	0.447	0.971	1.637	2.322	0	83	3.526
Hiv DNA	0.0014	0.953	5.017	0	0	9	
Human DNA	0.562	0.983	1.037	0.322	0	51	7
Loviana	0.509	0.958	1.027	1.069	0	77	5.574
Plasma	0.062	0.821	0.967	3.383	0.08	91	2.866
Primos	-0.01	0.044	3.55	0.594	0.5	78	1.295
Qp3	-0.008	0.605	0.96	-1.383	0.447	75	4.232
Sp500	0.015	0.859	1.026	3.569	0.035	82	2.585
Start	0.164	0.963	1.144	3.289	0.018	90	3.102
Tent	-0.243	0.971	1.029	0.477	0.5	13	3.401
Vanderpol	0.497	0.989	0.989	1.864	0.112	28	4.981

6 Resultados

A continuación se muestran los análisis correspondientes a tres de las propiedades más características para la cuantificación de la dificultad de predicción de las series tiempo. Para hacer la comparación se tomaron las clasificaciones de las series propuestas por Figueroa [3] (periódicas, cuasi-periódicas, caóticas, complejas estocásticas) la cual se basa en el comportamiento observado de la dinámica de señal.

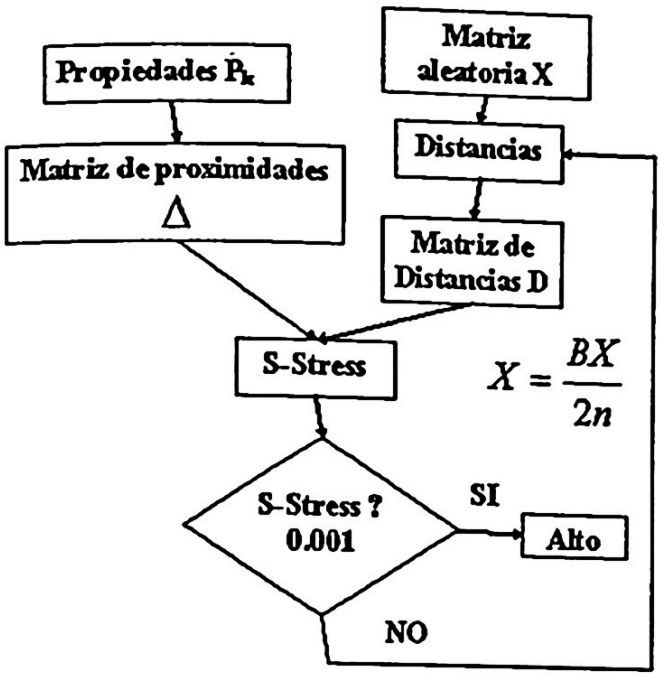

Diagrama 1. Algoritmo ASCAL

Al hacer el análisis MDS del exponente de Lyapunov (ver Fig. 1), se encontraron tres conjuntos que se diferencian por el nivel de inestabilidad que presenta la serie, cuando se le cambian las condiciones iniciales: las series del conjunto I tiene alta estabilidad a los cambios iniciales, las series del conjunto II son medianamente estables a cambios iniciales, y las series del conjunto III son poco estables a cambios iniciales.

Al hacer el análisis MDS del número de reglas de producción (ver Fig. 2), se encontraron tres conjuntos que se diferencian por el nivel de la complejidad computacional: las series del conjunto I tiene alta complejidad computacional, las series del conjunto II presentan complejidad computacional media, y las series del conjunto III presentan baja complejidad computacional.

El análisis MDS de entropía de Shannon (ver Fig. 3), nos muestra tres conjuntos que se diferencian por sus niveles de contenido de información: las series del conjunto I tiene alto contenido de información, las series del conjunto II presenta contenido de información medio, y las series del conjunto III tienen poco contenido de información.

Por último queremos hacer notar, que las relaciones de agrupamiento entre series de tiempo cambian dependiendo de la propiedad analizada con MDS, lo que nos indica que la clasificación (periódica, cuasi-periódica, caótica, compleja y estocástica) no permite caracterizar de forma completa su comportamiento. Por ejemplo, human DNA y HIV DNA aparecen en el mismo grupo en la Fig. 1 y en distintos grupos en las Fig. 2 y la Fig. 3. Ya que las dos series se toman como complejas se hubiera pensado que aparecería siempre en el mismo grupo con las diferentes propiedades.

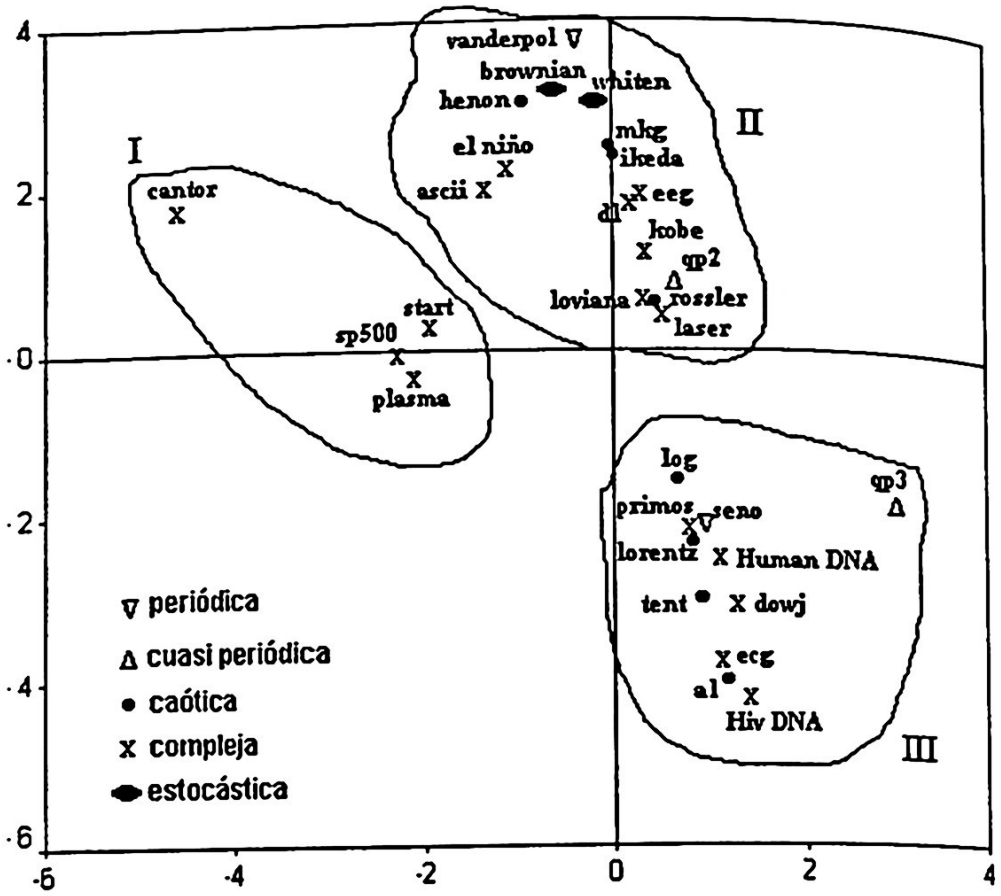

Fig. 1 MDS del exponente de Lyapunov de las treinta series

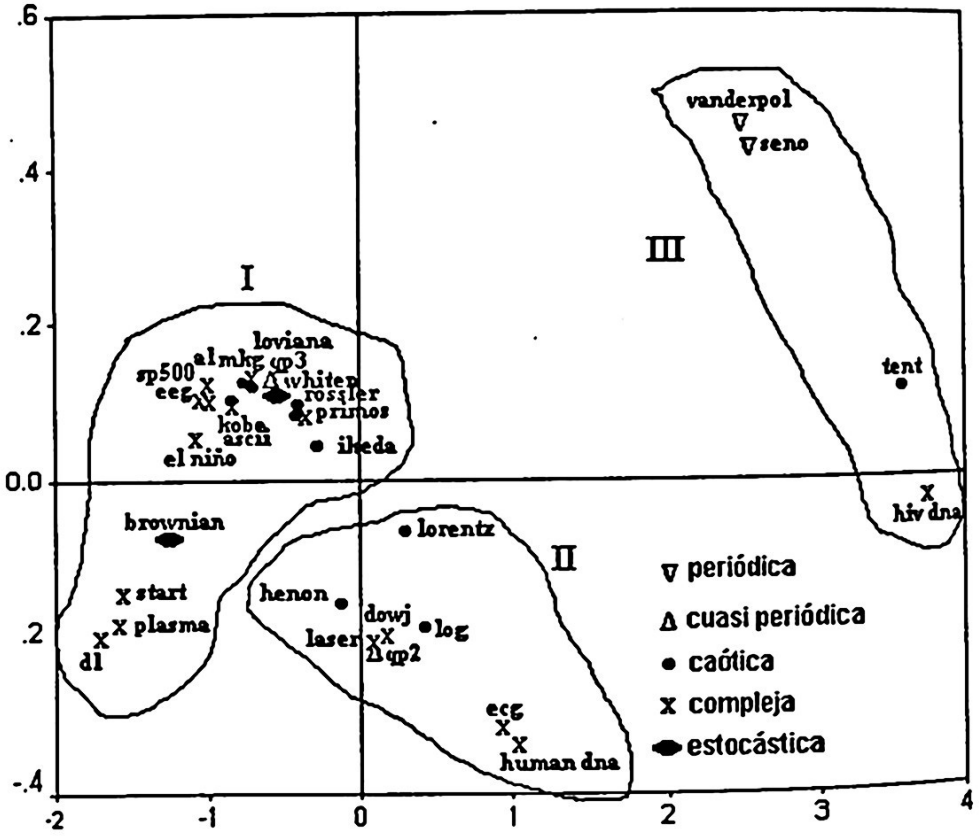

Fig. 2 MDS de reglas de producción de las treinta series

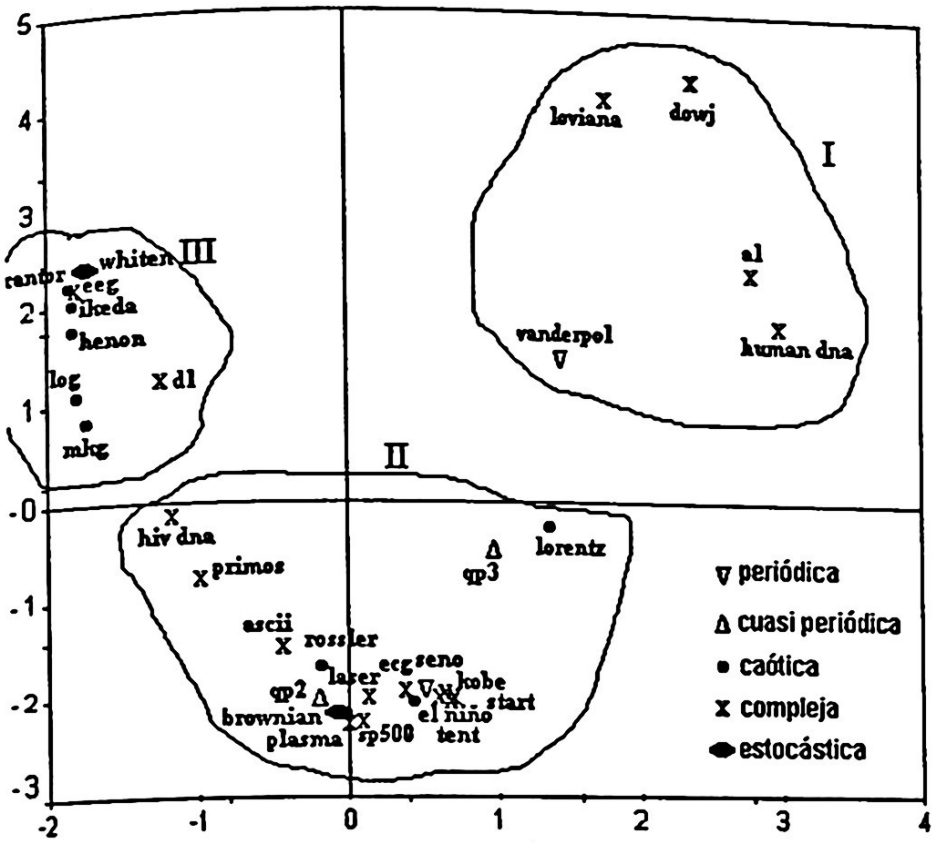

Fig. 3 MDS de entropía (Shannon) de las treinta series

7 Conclusiones

La relación entre la clasificación de referencia: (periódicas, cuasi-periódicas, caóticas, complejas y estocásticas) y las propiedades seleccionadas (exponente de Lyapunov, reglas de producción, entropía de Shannon) que permiten caracterizar la dificultad de predicción de las series de tiempo, es de carácter no lineal. Esto nos indica que clasificar una serie de tiempo únicamente por su comportamiento dinámico no es suficiente para caracterizar de forma completa a la misma. El análisis MDS ayuda a identificar clases o agrupaciones de series de tiempo con propiedades similares. La información del conjunto de propiedades puede ser utilizada por ejemplo, en la identificación de las relaciones entre las agrupaciones de series y así seleccionar el modelo de predicción más adecuado para las mismas.

Referencias

1. William R. Dillon & Matthew Goldstein (1984). "Multivariate Analysis: Methods and Applications" Editorial John-Wiley
2. Dallas E. Johnson (2000). "Métodos Multivariados Aplicados al Análisis de Datos" Thomson Editores
3. E. Bautista-Thompson, J. Figueroa-Nazuno. "Matriz de Conocimiento sobre la Complejidad de Predicción en Series de Tiempo". VII Congreso Iberoamericano

en Reconocimiento de Patrones. México, D. F. del 19 al 22 de Noviembre del 2002.

4. M.E Acevedo-Mosqueda, C.G. León-Vega & J. Figueroa-Nazuno "Medición de la Complejidad de Series de Tiempo". XLIV Congreso Nacional de Física, Morelia, Michoacán 2002
5. Espinosa-Contreras & J. Figueroa- Nazuno "Análisis del Comportamiento de la Perdida de Paquetes en la Red Internet con técnicas de la Dinámica No lineal". XLIV Congreso Nacional de Física, Morelia, Michoacán 2002
6. Robert Hilborn "Chaos and Nonlinear Dynamics An Introduction for Scientists and Engineers". Editorial Oxford University Press 2000
7. E. Bautista-Thompson, J. Figueroa-Nazuno. "Análisis de Propiedades que Caracterizan la Dificultad de Predicción en Series de Tiempo". XLV Congreso Nacional de Física. León, Guanajuato del 28 de Octubre al 1 de Noviembre del 2002.
8. R. Menchaca-Mendez, C. Sanchez-Rodrígez, J. Figueroa-Nazuno. "Predicción de Series de Tiempo Mediante Análisis Gramatical". XLIII Congreso Nacional de Física. Puebla, Puebla.